Courrier des lecteurs

À propos du maximum d'une variance

À la suite de l'article de Serge Chesney paru dans le bulletin 443 page 740, je voudrais proposer une autre démonstration de la recherche du maximum de la variance d'une série statistique bornée de moyenne fixée. Cette démonstration, moins élégante, n'utilise pas la géométrie de $\mathbf{R}^n$, mais peut être comprise par un élève de terminale.

I Rappel du problème

Comment choisir n nombres x_1, x_2, …, x_n appartenant à l'intervalle $[0 ; a]$ de moyenne m tels que la variance de la série (x_i) soit maximale ?

II Recherche intuitive

Pour que la variance soit maximale, il suffit que les (x_i) s'écartent le plus possible de la moyenne m.
On peut alors penser que la variance est maximale lorsque les valeurs égales aux extrémités de l'intervalle $[0 ; a]$ sont les plus nombreuses possibles. On se pose alors la question suivante : peut-on trouver n nombres égaux à 0 ou a de moyenne m (ou de somme nm) ? La réponse est non sauf cas particulier. On se pose alors le problème suivant : Peut-on trouver n nombres de $[0 ; a]$ de somme nm tels qu'au moins $n - 1$ de ces nombres soient égaux à 0 ou a ? Plus précisément existe-t-il deux entiers p et q et un réel $r \in [0 ; a]$ tels que $p \times 0 + q \times a + r = nm$ et $p + q = n - 1$? p sera le nombre de valeurs égales à 0, q le nombre de valeurs égales à a et r la valeur éventuellement différente de 0 et de a.
Si $nm \leq a$, on prend $q = 0$, $p = n - 1$, $r = nm$.
Si $a < nm \leq 2a$, on prend $q = 1$, $p = n - 2$, $r = nm - a$.
....
Si $ka < nm \leq (k + 1)a$, on prend $q = k$, $p = n - 1 - k$, $r = nm - ka$.
....
Si $(n - 1)a < nm \leq na$, on prend $q = n - 1$, $p = 0$, $r = nm - (n - 1)a$.

Comme $nm \in [0 ; na]$, le problème a toujours une solution que l'on note $y_1, y_2, \ldots, y_n$.
Il reste à montrer que cette solution maximise la variance, c'est-à-dire que, quels que soient les nombres x_1, x_2, …, x_n appartenant à $[0 ; a]$ et ayant pour moyenne m,

$$\frac{1}{n}\sum_{i=1}^{n}(m - x_i)^2 \leq \frac{1}{n}\sum_{i=1}^{n}(m - y_i)^2.$$

Pour cela on va utiliser le théorème de Koenig-Huygens.

III Théorème de Koenig-Huygens

Soient $x_1, x_2, \ldots, x_n$ n nombres de moyenne m, et m' un réel quelconque :

$$\frac{1}{n}\sum_{i=1}^{n}(m-x_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(m'-x_i)^2 - (m-m')^2.$$

Démonstration :

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(m-x_i)^2 &= \frac{1}{n}\sum_{i=1}^{n}(m-m'+m'-x_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(m-m')^2 + \frac{1}{n}\sum_{i=1}^{n}(m'-x_i)^2 + \frac{2}{n}\sum_{i=1}^{n}(m-m')(m'-x_i) \\ &= (m-m')^2 + \frac{1}{n}\sum_{i=1}^{n}(m'-x_i)^2 + \frac{2}{n}(m-m')(nm'-nm) \\ &= \frac{1}{n}\sum_{i=1}^{n}(m'-x_i)^2 - (m-m')^2.\end{aligned}$$

On en déduit que pour comparer les variances de deux séries de même moyenne, il suffit de comparer la moyenne (ou la somme) des carrés des écarts par rapport à un nombre m' convenablement choisi.

IV Démonstration

On veut montrer que si $ka < nm \leq (k + 1)a$, où k est un entier compris entre 0 et $n - 1$, la variance de toute série $(x_1, x_2, \ldots, x_n)$ de moyenne m est inférieure ou égale à celle de la série $(0, 0, \ldots, 0, nm - ka, a, a, \ldots, a)$ dans laquelle il y a k valeurs égales à a et $n - k - 1$ valeurs égales à 0. Il suffit pour cela, d'après le théorème de Koenig-Huygens, de montrer que :

$$\frac{1}{n}\sum_{i=1}^{n}(m'-x_i)^2 \leq (n-k-1)(m'-0)^2 + (m'-(nm-ka))^2 + k(m'-a)^2$$

où m' peut être choisi arbitrairement. Distinguons deux cas pour le choix de m'.

<u>Premier cas :</u> Si le nombre de x_i appartenant à l'intervalle $[nm - ka\ ;\ a]$ est supérieur à k, on prend pour m' le centre de l'intervalle $[nm - ka\ ;\ a]$, c'est à dire : $m' = \dfrac{nm-ka+a}{2}$. Comme $nm - ka \in [0\ ;\ a]$, on a : $m' \in \left[\dfrac{a}{2}\ ;\ a\right]$. On choisit alors k valeurs x_i de l'intervalle $[nm - ka\ ;\ a]$. Pour chacune de ces k valeurs, on a : $(m' - x_i)^2 \leq (m' - a)^2$ puisque a est une extrémité de l'intervalle et m' le centre de l'intervalle. Il y a au moins une autre valeur x_i dans $[nm - ka\ ;\ a]$. Pour cette valeur, on a : $(m' - x_i)^2 \leq (m' - (nm - ka))^2$ car $nm - ka$ est une extrémité de l'intervalle. Pour les $(n - k - 1)$ autres valeurs on a : $(m' - x_i)^2 \leq (m' - 0)^2$. En effet si $x_i \geq nm - ka$, $|m'-x_i| \leq \dfrac{a}{2}$ et $m'-0 \geq \dfrac{a}{2}$ et si $x_i < nm - ka$, alors

$0 < m' - x_i \leq m' - 0$. On a donc :

$$\sum_{i=1}^{n}(m'-x_i)^2 \leq k(m'-a)^2 + (m'-(nm-ka))^2 + (n-k-1)(m'-0)^2.$$

Deuxième cas : Si le nombre de x_i appartenant à l'intervalle $[nm - ka ; a]$ est inférieur ou égal à à k, alors le nombre de x_i appartenant à l'intervalle $[0 ; nm - ka]$ est supérieur ou égal à $n - k$ et on prend pour m' le centre de l'intervalle $[0 ; nm - ka]$. On choisit alors $(n - k - 1)$ valeurs x_i dans cet intervalle. Pour chacune de ces valeurs on a : $(m' - x_i)^2 \leq (m' - 0)^2$. Il y a au moins une autre valeur x_i dans $[0 ; nm - ka]$. Pour cette valeur, $(m' - x_i)^2 \leq (m' - (nm - ka))^2$. Pour les k autres valeurs, on a : $(m' - x_i)^2 \leq (m' - a)^2$ car si $x_i \leq nm - ka$, on a $|m'-x_i| \leq \frac{a}{2}$ et $a - m' \geq \frac{a}{2}$ puisque $m' \leq \frac{a}{2}$ et si $x_i > nm - ka$, alors $0 < x_i - m' \leq a - m'$. D'où :

$$\sum_{i=1}^{n}(m'-x_i)^2 \leq k(m'-a)^2 + (m'-(nm-ka))^2 + (n-k-1)(m'-0)^2.$$

Donc le maximum de la variance est la variance de la série $(0, 0, \ldots, 0, nm - ka, a, a, \ldots, a)$ c'est à dire :

$$\frac{1}{n}[(nm-ka)^2 + ka^2] - m^2 = \frac{1}{n}[n(n-1)m^2 - 2nmka + k(k+1)a^2]$$

où k est la partie entière de $\frac{nm}{a}$.

Pierre Carriquiry